ANNUAL REVIEWS

Annual Review of Ecology, Evolution, and Systematics

The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics

Adam D. Leaché[1] and Jamie R. Oaks[2]

[1]Department of Biology and Burke Museum of Natural History and Culture, University of Washington, Seattle, Washington 98195; email: leache@uw.edu

[2]Department of Biological Sciences, Auburn University, Auburn, Alabama 36849; email: joaks@auburn.edu

Annu. Rev. Ecol. Evol. Syst. 2017. 48:69–84

The *Annual Review of Ecology, Evolution, and Systematics* is online at ecolsys.annualreviews.org

https://doi.org/10.1146/annurev-ecolsys-110316-022645

Keywords

phylogenetic inference, single nucleotide polymorphism, molecular systematics, phylogeography, RADseq, genealogy of life

Abstract

Resolving the genealogy of life—the phylogenetic relationships that describe the evolutionary history of species—remains one of the great challenges of systematic biology. The recent proliferation of DNA sequencing technologies has sparked a rapid increase in the volume of genetic data being applied to phylogenetic studies. Single nucleotide polymorphism (SNP) data, ubiquitous genetic markers once considered reserved for population genetic studies, are now being applied in phylogenetics research at deep evolutionary timescales. The potential for SNPs to resolve contentious phylogenetic problems while researchers also investigate population demographics is promising, yet serious challenges remain with respect to data collection, assembly, modeling, and analysis. The low cost and ease of collecting SNPs suggest that they will remain an important source of genetic information for inferring phylogenies across time periods ranging from the Anthropocene to the Cretaceous.

1. INTRODUCTION

Single nucleotide polymorphism (SNP): a single base-pair difference among chromosome sequences

Orthologous: SNPs that are related through shared common ancestry

Restriction site associated DNA sequencing (RADseq): a method to discover and screen SNPs in a reproducible subsection of the genome, yielding thousands of SNPs per individual

A single nucleotide polymorphism (SNP) is an orthologous nucleotide position that is variable across the genomes under study. SNPs are the result of mutations that produce base-pair differences among chromosome sequences. Their high abundance and genome-wide distribution make them a valuable source of genetic variation for studies of population demography, adaptation, and genome evolution (Brumfield et al. 2003, Morin et al. 2004). Recently, SNPs have started to play an increasingly important role in phylogeographic and phylogenetic studies at much deeper evolutionary timescales. Economical solutions for collecting comparative genomic data are always at a premium in molecular systematics, and in this regard SNP data are a practical choice compared with alternative methods that are more time-consuming and expensive.

Empirical phylogenetic studies using SNPs have become commonplace across diverse taxa and studies ranging from vertebrate adaptive radiations (Wagner et al. 2013), recent insect diversification (Emerson et al. 2010), recalcitrant phylogenetic relationships (Eaton & Ree 2013, Herrera & Shank 2016), and phylogeographic structure of pathogens responsible for infectious disease in humans (Filliol et al. 2006). The exponential growth of such studies over the past two decades (**Figure 1**) is a clear indication that the scientific community is interested in using SNPs for phylogenetic inference and that the number of such studies will continue to grow.

New data collection techniques are driving the proliferation of SNPs in evolutionary and systematic studies. Methods for the de novo sequencing and identification of SNPs, such as restriction site associated DNA sequencing (RADseq) and genotyping by sequencing (GBS) methods, are increasingly popular in ecology and evolution (Baird et al. 2008, Seeb et al. 2011, Ree & Hipp 2015). Key attributes that make these methods appealing include their low cost, ease of implementation, applicability to nonmodel species, and scalability to large sample sizes. These factors have helped

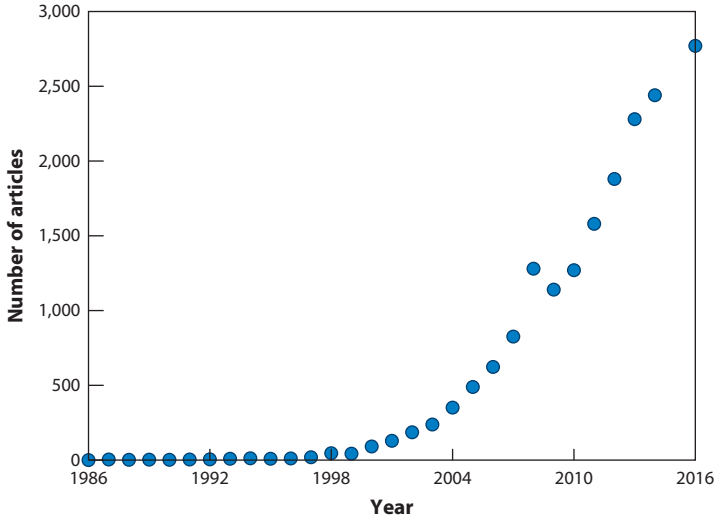

Figure 1

Increase in the use of single nucleotide polymorphism (SNP) data in phylogenetics as indicated by the number of articles published annually. The citation search was conducted using Google Scholar on January 10, 2017, using the exact search phrases "SNP" and "phylogeny" and with at least one of the words "evolution" or "systematics." Searches were run for each year individually for the past 30 years.

researchers acquire SNPs, and thus scale up to genome-wide data, for studies that were typically conducted using conventional DNA sequencing (Schuster 2007).

What differentiates SNP phylogenetics from traditional approaches that use gene sequences? Although SNPs are not fundamentally different from sequence data—they share the same basic unit, the nucleotide—the manner in which data are collected has important ramifications for data analysis. DNA sequences are linear strings of nucleic acids composed of two general types of characters once they are aligned to one another: constant and variable characters. SNP data acquisition typically circumvents, either biochemically or bioinformatically, the sequencing of constant characters. This acquisition bias leads to data composed exclusively of variable characters (SNPs), either as a direct result of the data collection strategy, or after bioinformatics steps selectively remove the constant characters. Acquisition bias is problematic because it can lead to extreme overestimation of evolutionary rates and inaccurate topologies, branching times, and support values (Lewis 2001, Leaché et al. 2015a).

Acquisition bias: distorting the characteristics of data during collection or postprocessing

Coverage: the number of reads that include a given nucleotide in the reconstructed sequence

What does the future hold for the utility of SNP data in phylogeographic and phylogenetic analysis? What are the advantages and disadvantages of SNPs versus traditional DNA sequence data? Do empirical and simulation studies agree on how broadly SNPs can be applied across the genealogy of life? How do new methods for estimating SNP phylogenies compare, and what are the circumstances that make one method more suitable compared with another? Our goal is to review the relevant literature and discuss where the field of phylogenetics currently stands on these open questions.

2. ADVANTAGES OF SNP PHYLOGENETICS

Phylogenetic questions are not limited to specific taxonomic ranks or geological periods of time. They span a continuum from recent divergences at the phylogeographic level (Hickerson et al. 2010, Edwards et al. 2016a) to ancient diversification among the major lineages of life (Hinchliff et al. 2015, Hug et al. 2016). Therefore, the importance of selecting the most appropriate type of data for addressing a specific phylogenetic question has remained at the forefront of study design issues in systematics (Harris et al. 2014, Streicher et al. 2016). The current use of SNPs for widely diverse research questions suggests that they are suitable for addressing a wide range of questions across broad evolutionary timescales.

The ability to quickly and easily collect large numbers of SNPs without a significant investment in genomic resources is appealing to researchers studying nonmodel species (Nieto-Montes de Oca et al. 2017). RADseq methods in particular are among the most cost-effective approaches available (Ree & Hipp 2015). Although costs are dropping, sequencing whole genomes remains prohibitively expensive for studies of nonmodel systems. RADseq circumvents this problem by sampling small but widely distributed genomic sequences. The general protocol of a RADseq experiment involves restriction enzyme digestion of a genome, followed by size selection, adapter and barcode ligation, and sequencing. Advantages of RADseq data include the wide dispersion of SNPs across chromosomes, affordability and ease of data collection, high read overlap resulting in deeper sequencing coverage, and the overall high information content (Harvey et al. 2016). Other popular methods for collecting phylogenomic data typically require either transcriptome data (Portik et al. 2016) or genome alignments for developing probe sequences (Faircloth et al. 2012), either of which adds substantial cost and time to a project.

A long-standing tradition in the field of phylogeography is the determination of the phylogenetic relationships among distinctive populations within species (Avise 2000). Gene sequences with high variability and rapid coalescence times are well suited for this purpose (Zink & Barrowclough 2008). However, accurate phylogeographic inference requires thorough analyses of

multiple independent genetic markers because any single locus is only a single sample from highly stochastic ancestral (or genealogical) processes (Degnan & Rosenberg 2009). The accuracy of important population parameters, including population size, divergence times, and migration rates, increases when more loci are sampled (Felsenstein 2006). Phylogeographic studies have therefore capitalized on the large numbers of loci afforded by SNPs to estimate population histories in fine detail (Reitzel et al. 2013).

Sequence capture: a technique for isolating and sequencing preselected regions of the genome, typically with the aid of probe sequences that complement the genes of interest

Linkage: the inheritance of SNPs owing to their close proximity to one another on a chromosome

Allelic dropout (ADO): biased sampling of alleles because of mutations at enzyme recognition sequences

Phylogeographic studies using SNPs typically contain more loci compared with studies using traditional Sanger sequencing or sequence capture protocols. An evaluation of published phylogeographic data sets showed a sharp increase in the number of SNPs per study starting in 2013 and predicted that the median number of SNPs per data set for studies published by the end of 2016 would approach 20,000 (Garrick et al. 2015). In addition to providing more data for resolving difficult and/or recent phylogenetic relationships, these larger data sets also provide greater precision for population genetic parameter estimates (Harvey et al. 2016). Another benefit of SNPs is the fact that they can help alleviate the concern of intragenic recombination—given that a SNP is a single nucleotide, there is no possibility for intragenic recombination. Traditional phylogenetic methods using gene sequences typically assume that the alignment is free from recombination and that each site evolved along a single, shared genealogy. However, recombination can produce a mosaic of different genealogies across the sites of a sequence alignment, and precautions must be taken to prevent alignments from containing recombination break points, which can mislead results (Posada & Crandall 2002). This does not free SNPs from the assumption that they are either on different chromosomes or far enough apart on the same chromosome that their genealogies are effectively independent (i.e., any linked SNPs must still be phased). Many studies are confronting the issue of linkage by mapping SNPs to full genomes (Kawakami et al. 2014, Sutherland et al. 2016).

3. SNP DATA STRUCTURE

3.1. Hierarchical Missing Data

Collecting orthologous SNPs from across distantly related species can be a challenge, and the problem becomes more pervasive at deeper evolutionary timescales. As a consequence, a pattern that has emerged from phylogenetic studies utilizing SNPs is that the amount of missing data increases with the number of loci and the sample size (Wagner et al. 2013). This missing data problem is a result of the data collection procedure; methods that rely on the conservation of restriction enzyme cut sites to isolate and sequence orthologous DNA fragments (such as RADseq) will fail to sample a locus if a mutation occurs in the enzyme recognition site. This problem, known as allelic dropout (ADO) (Arnold et al. 2013), predicts that fewer loci will be recovered that overlap across more distantly related species (**Figure 2**). The ADO problem is especially relevant to RADseq methods because the presence of polymorphism within the restriction site will make it impossible to observe the associated SNP allele in all samples. The opposite problem also occurs, where nontarget loci are unintentionally processed and sequenced owing to nonspecific restriction enzyme cutting (star activity). An empirical study of the zebra finch (*Taeniopygia guttata*) found that thousands of extra loci resulted from star activity, although their low sequencing depth precluded large negative consequences on subsequent phylogenetic inference (DaCosta & Sorenson 2014).

The problem of ADO can occur across taxonomic scales ranging from populations, to species, to clades, because mutations in restriction enzyme recognition sequences will be inherited. For example, by coding missing data as a presence-absence polymorphism, DaCosta & Sorenson (2016) demonstrated that the pattern of missing data in RADseq data contains phylogenetic signals. This

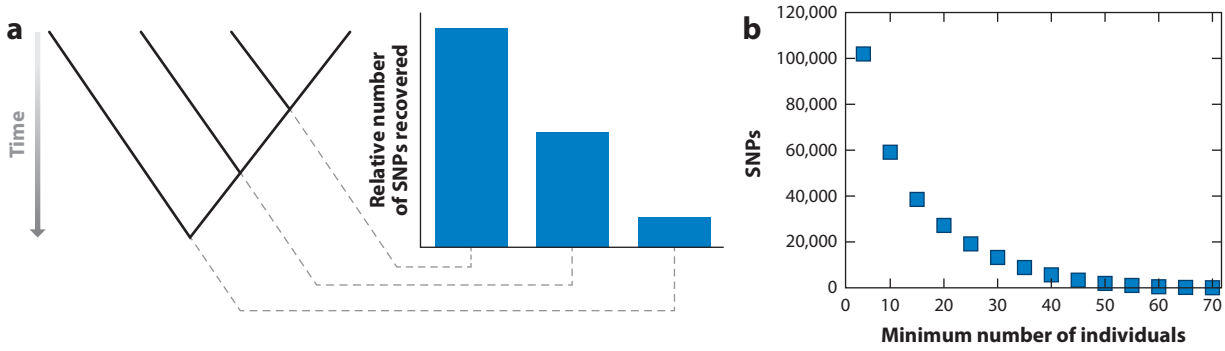

Figure 2

Characteristics of single nucleotide polymorphism (SNP) data quantities derived from restriction site associated DNA sequencing (RADseq) methods. (*a*) Allelic dropout causes a sharp decrease in the recovery of orthologous SNPs between distant relatives. The phylogenetic tree characterizes a typical pattern encountered with RADseq data collection methods: Close relatives tend to share many SNPs, whereas far fewer are shared among more distant relatives. Deciding the amount of acceptable missing data at a locus is an important aspect of SNP analyses. In general, practitioners can modulate the amount of missing data in a data matrix by applying a penalty to the minimum number of individuals that are required to have data at a locus in order to include that locus in a final data matrix. (*b*) The number of SNPs recovered declines drastically as more individuals are required to have data at each locus. As the minimum number of individuals is decreased, the penalty against missing data is relaxed, and the number of SNPs increases exponentially. This pattern is typical of real data (Wagner et al. 2013). The data presented here are adapted from empirical data for phrynosomatid lizards (Leaché et al. 2015a).

genealogical bias in the recovery of loci has important ramifications for population-level and phylogenetic inferences.

At the population level, ADO will produce biased estimates of common summary statistics [i.e., π, θ, Tajima's D, and F_{ST} (Arnold et al. 2013)]. The frequency of ADO has been derived analytically (Gautier et al. 2013); ADO within a population depends on the mutation rate per nucleotide scaled by the effective population size parameter, theta ($\theta = 4\text{Ne}\mu$). Simulation work has shown that ADO will overestimate genetic variation both within and between populations and that the bias is most severe for large effective population sizes (Gautier et al. 2013). This counterintuitive result is explained by the fact that ADO tends to remove a portion of high-frequency alleles, which causes an increase in the minor allele frequency (MAF) and thus an overall inflation of heterozygosity (Gautier et al. 2013).

Promising new data collection methods are available that appear to be more robust to ADO (Ali et al. 2016, Hoffberg et al. 2016, Suchan et al. 2016). These new methods are best classified as hybrid approaches that capitalize on the advantages of both RADseq and sequence capture. In general, these methods begin by developing a library of RAD loci using one or more high-quality samples and then design baits or probes to target specific loci in subsequent hybridization enrichment procedures. The result is a data collection strategy that still targets large numbers of RAD loci but circumvents ADO by targeting the sequence of the entire locus instead of a short restriction enzyme recognition sequence. One empirical study of a clade of plants endemic to the European Alpine system (*Primula* sect. *Auricula*) reported a 20% decrease in missing data using this approach (Boucher et al. 2016). These methods also have the potential to enable researchers to begin gathering genomic sequence data from museum specimens (Yeates et al. 2016, Linck et al. 2017), where natural degradation of DNA sequences precludes standard library preparation procedures (Hykin et al. 2015).

Theta (θ): a measure of genetic diversity equivalent to four times the product of the effective population size and the mutation rate per site per generation ($4\text{Ne}\mu$)

Minor allele frequency (MAF): the frequency of the second most common allele at a locus in a given population

3.2. Random Missing Data

It is possible that the prevalence of ADO is overstated in the literature by empirical studies and that it is not the leading source of missing data. One empirical study found that sequencing effort had a more significant impact on patterns of missing data than ADO (Eaton et al. 2015). A subsequent comparison of 10 empirical data sets supported the idea that insufficient or uneven sequencing coverage can account for similar levels of missing data as ADO (Eaton et al. 2016).

Ascertainment bias: analysis bias from subjective decisions made during data sampling, such as the analysis of only the most variable loci

Gene tree: a genealogical tree of sequences from a genomic region

Gene tree heterogeneity: topological and branch length variability across genetic loci that results from natural biological processes

3.3. Ascertainment Bias

SNP data collection techniques that rely on probe sequences targeting regions found to vary in an initial survey of a small number of samples (a panel) can suffer problems related to ascertainment bias, as any variants not present in this initial survey are typically missed (Kuhner et al. 2000, Nielsen 2000, McGill et al. 2013). This panel approach will create a bias toward SNPs that have alleles at intermediate frequencies. All downstream analyses will assume that the SNPs are a random sample of variable sites across the genome. So, any result that is supported by patterns of alleles at intermediate frequency would be suspect. Ascertainment bias can affect the inference of admixture proportions and population histories, both of which are important aspects of phylogeographic inference (McTavish & Hillis 2015). Some phylogeographic studies have tried to minimize ascertainment bias by using RADseq methods to discover genome-wide variation across a wide panel of samples and then select a subset of SNPs of interest for subsequent genotyping assays that can process thousands of samples. This approach is particularly appealing in threatened or endangered species, where reproducible SNP genotyping of large numbers of samples can help monitor and manage species (Shafer et al. 2015, Garner et al. 2016, Stetz et al. 2016).

4. THE PHYLOGENETIC UTILITY OF SNPS

4.1. Computer Simulation Studies

Several computer simulation studies have evaluated whether SNPs can provide accurate and robust phylogenetic results for groups as old as 60 million years (Rubin et al. 2012, Cariou et al. 2013). These studies obtained RADseq data from mammals, *Drosophila*, and yeast using an in silico method, whereby data are harvested from sequenced genomes using software processing pipelines that mimic the steps of the molecular lab procedures. One study used published genomes of 23 primates to conduct a standardized comparison of four phylogenomic methods, including several RADseq techniques and sequence capture (Collins & Hrbek 2015). The study found that most data conflicts were associated with a difficult clade comprising a rapid diversification and emphasized the importance of accurately modeling heterogeneous data, such as gene tree heterogeneity. The ability to reconstruct the known phylogenies for these groups sets an important benchmark regarding the expected accuracy of SNPs in groups that have not been previously studied and for which there is no existing phylogeny for comparison.

The in silico restriction digest experiments using real genome data are useful, but there is an advantage to conducting simulations where the true tree and branch lengths are known. Simulations of contrived data under fabricated evolutionary histories are useful for quantifying potential biases. The few simulation studies conducted to date have challenged our conventional wisdom regarding issues related to the effects of missing data on phylogenetic accuracy.

ADO can bias estimates of population genetic parameters at shallow scales (Arnold et al. 2013), which suggests that phylogenetic inference should also be sensitive. A simulation study by Huang

& Knowles (2016) showed that ADO has effects other than reducing the amount of data available for analysis. Specifically, loci with high mutation rates were the most likely to be missing, and the authors therefore cautioned against removing data, which could result in biasing the data in favor of loci with low mutation rates. However, loci with large amounts of missing data tend to produce discordant topologies with increased branch length errors (Leaché et al. 2015a). More work is needed to develop objective methods for evaluating and selecting missing data thresholds that avoid negative consequences for phylogenetic analysis.

Ultraconserved element (UCE): a highly conserved region of organismal genomes shared among evolutionarily distant taxa

Species tree: a phylogenetic tree for a set of species that underlies the gene trees at individual loci

Parsimony-informative: a variant that is shared by two or more samples

4.2. Empirical Studies

Comparative studies of SNPs versus gene sequences collected for the same set of samples provide a powerful approach for directly comparing and evaluating phylogenetic performance of different types of data. Though few studies of this nature are currently available, those that exist span a fairly wide range of divergence times and therefore are helpful in understanding the consequences of marker choice on phylogenetic inference. They show that there is typically large disparity in the number of SNPs obtained versus gene sequences, and even when all segregating sites are mined from the gene sequences, the SNPs still outnumber the gene sequence variation by orders of magnitude (Harvey et al. 2016). However, several major obstacles to comparing gene sequences with SNPs exist, including assembling equivalent data sets for balanced comparisons, identifying phylogenetic methods that can handle both types of data without making simplifying assumptions that can bias the results, and identifying how to best measure accuracy and performance. In general, comparative studies have concluded that SNPs and gene sequences produce broadly concordant topological results. However, the results can be sensitive to the assumptions used during data assembly (Leaché et al. 2013, Harvey et al. 2016, Manthey et al. 2016).

Harvey et al. (2016) investigated the utility of SNPs versus gene sequences among populations of birds (genus *Xenops*) that began diversifying approximately 5 million years ago. The data were collected using RADseq for SNP loci and sequence capture of ultraconserved elements (UCEs) for gene sequences (Smith et al. 2014). The SNP loci outnumbered the gene sequences by two orders of magnitude (SNP loci = 158,329; gene sequences = 1,358). Despite drastically different data set attributes, the estimates of genetic distances, population sizes, and phylogenetic trees were quite similar.

Another avian study compared a clade of 11 species in the bird genus *Piranga* that diverged approximately 6 million years ago (Manthey et al. 2016). This study found that SNPs (1,128) and gene sequences (UCEs; 189 loci) supported the same species level phylogeny. An interesting aspect of this study was the decision to compare species trees estimated using only parsimony-informative SNPs (singleton SNPs were removed) and UCEs containing at least 10 parsimony-informative sites. Phylogenetic analyses of SNPs that only include parsimony-informative sites may produce reliable topologies (but see Allman et al. 2010), but removing certain sites (singletons) should bias other parameters of the species tree, including population sizes and divergence times. A comparative study of phylogenetic relationships estimated using sequence capture data and SNPs for a clade of North American phrynosomatid lizards included divergences as old as 50 million years (Leaché et al. 2015b). The study found congruent phylogenetic relationships when comparing the largest data sets (584 gene sequences versus 2,670 SNPs); however, the phylogeny estimated from the SNP data was sensitive to the thresholds used for locus assembly, paralog detection, and missing data. The conflicts among the SNP trees were restricted to short internal branches of the phylogeny. This observation matches the conclusions of Collins & Hrbek (2015) and further underscores the

importance of continuing to explore the relationships between data assembly assumptions and phylogenetic inference, particularly when addressing difficult phylogenetic questions.

Concatenation: joining independent loci together to assemble one large supergene

5. METHODS FOR ESTIMATING SNP PHYLOGENIES

5.1. Concatenation

Data concatenation entails stitching SNPs together end-to-end to construct one exceptionally long super locus (**Figure 3**). The process of data concatenation does not produce a phylogeny.

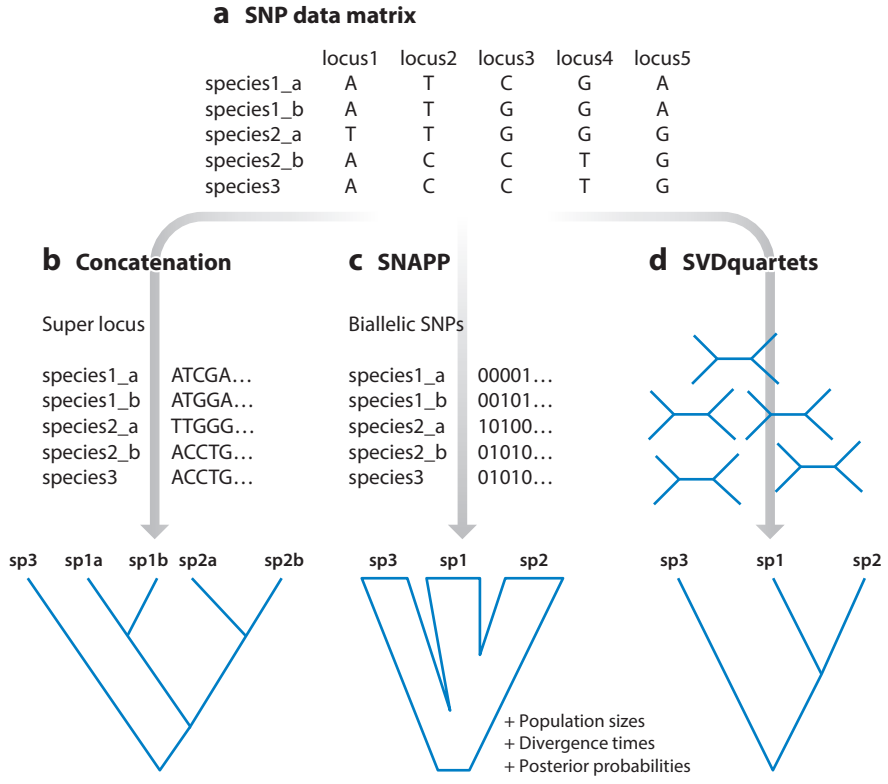

Figure 3

An overview of single nucleotide polymorphism (SNP)-based phylogenetic methods, including concatenation, SNAPP (SNP and amplified fragment length polymorphism phylogenies), and SVDquartets (singular value decomposition for quartets). (*a*) A SNP data matrix containing one SNP per locus for three species; two of the species (species1 and species2) are represented by multiple individuals. (*b*) Concatenation stitches independent loci together, and as a result, the outcome from a phylogenetic analysis will resemble a typical gene tree where the samples for each species are represented as tips in the phylogeny. This process violates the assumption that different genes have independent histories. (*c*) Coalescent-based species tree methods, including SNAPP, account for incomplete lineage sorting and provide phylogenetic trees with species at the tips and not individual samples. Coalescent methods also provide population size and divergence time estimates, as well as posterior probabilities for species trees. (*d*) The SVDquartets method estimates a species tree by exhaustively sampling all combinations of four taxa from the data matrix (or, heuristically, for exceptionally large numbers of samples), inferring a tree for each quartet, and then using a quartet assembly algorithm to combine all sampled quartets into a species tree.

Instead, the concatenated data are analyzed using a variety of phylogenetic inference approaches, including distance methods, maximum parsimony, maximum likelihood, or Bayesian inference.

Concatenation remains a commonly used method for estimating SNP-based phylogenetic trees, although the method suffers from many shortcomings. Concatenation is problematic because it ignores incomplete lineage sorting, and in doing so it assumes that all SNPs share the same coalescent history. This simplifying assumption is unjustifiable (Edwards et al. 2016b). A fundamental advantage of and motivation for collecting multilocus data (including SNPs) is to obtain independent genealogical samples for measuring important population parameters, including divergence times, population sizes, and species relationships (Edwards & Beerli 2000, Rannala & Yang 2003), and concatenation essentially reduces the number of independent samples to one. This massive reduction in variance can lead to serious problems, including overestimated support for phylogenetic relationships (false precision), and a bias toward incorrect trees (low accuracy) (Liu et al. 2015, Xu & Yang 2016). When estimating phylogenetic trees, it is better to have an accurate result that is imprecise compared with a precise result that is inaccurate (Swofford et al. 2001).

However, when concatenation is necessary, studies suggest it is better to apply this method to the original sequence alignments from which the SNPs were extracted. Removing constant sites results in acquisition bias that can inflate branch length estimates and, in extreme cases, produce an inaccurate phylogeny (Leaché et al. 2015a). Providing the analysis with the empirical frequencies of constant sites requires trivial additional computation time and can help alleviate some significant branch length estimation biases (Leaché et al. 2015a).

5.2. Bypassing Gene Trees

New methods are available for inferring species trees without the requirement of first (or simultaneously) estimating gene trees for each locus. The methods implemented in SNAPP (SNP and amplified fragment length polymorphism phylogenies) (Bryant et al. 2012) and PoMo (polymorphism-aware phylogenetic model) (De Maio et al. 2013, 2015a; Schrempf et al. 2016) offer innovative full-likelihood approaches to inferring species trees directly from SNP data. In addition to some of the benefits common to all SNP-based methods mentioned above (no phasing of haplotypes and no intralocus recombination), both of these methods offer a very important computational benefit: They analytically integrate gene trees out of the equation. As a result, these methods do not have to numerically sample all the parameters associated with the genealogies of all loci, many of which are strongly correlated. For other methods that rely on gene sequences and gene trees, the large number of correlated parameters greatly limits the number of loci that can be analyzed (but see Chung & Hey 2017). However, bypassing gene trees allows SNAPP and PoMo to reconstruct moderate-sized species trees (approximately 20 species) from very large numbers of SNPs (thousands of loci) in reasonable time (tens of hours).

SNAPP and PoMo both modify the state space of nucleotide substitution models from single nucleotides (i.e., A, C, G, or T) to encompass sets of alleles or allele counts (e.g., 7 As and 3 Ts). This allows both methods to model state changes due to mutations and drift. SNAPP uses a reverse-time coalescent model, which allows it to use the number of sampled alleles within each population to constrain the possible allele counts for ancestral populations. However, the number of possible allele counts can get very large toward the base of trees with many taxa, because all the alleles sampled at the tips that descend from the branch must be included (even if the probability that none of the alleles have coalesced this deep in the tree is small, it is still not zero). PoMo uses a forward-time Moran model to derive the probabilities of state changes due to drift. This requires the state space of possible allele counts to encompass a whole population of alleles (not just

the sampled alleles). Although this virtual population of alleles can be quite small (≈10; Zeng & Charlesworth 2009, Schrempf et al. 2016), it can still adequately explain the site patterns generated from a much larger biological population.

Quartet: an unrooted tree for four taxa

In contrast, the number of possible allele counts in the virtual population is very large. Unlike the SNAPP model, however, it remains the same size for all branches of the tree. Nonetheless, both methods use a large matrix to describe the rates of change among the different allele-count states due to either mutation or drift. By exponentiating these large matrices, they can determine the probability of change between any two allele-count states across a branch of a given length, while effectively integrating over all possible genealogies and mutational histories.

SNAPP and PoMo make some limiting assumptions to reduce the number of possible allele counts that need to be considered. Currently, the SNAPP model only accommodates biallelic characters (**Figure 3**). This introduces some arbitrariness about how to reduce SNPs from four states to two states. If the rates of mutation between the two states are assumed to be equal, this is not a problem. However, if the mutation rates are allowed to vary, then the likelihood is sensitive to how the nucleotides at each site are coded as binary. Furthermore, any sites with more than two nucleotides observed are not allowed. In contrast, the PoMo model accommodates all four nucleotides. But to do so, it must assume that drift is fast enough (the effective population size is small enough) relative to the mutation rates such that a second mutation at the same site does not occur prior to the fixation or loss of the previous mutation. As a result, any sites for which there are more than two nucleotides within a species are not allowed. These will obviously be much rarer than sites that have more than two nucleotides across all species, which are not allowed in SNAPP's biallelic model.

Although both models can condition on collecting only variable characters, neither method corrects for discarding sites with two or more nucleotides within a species (PoMo) or across species (SNAPP). This can introduce an additional source of bias. Another difference between the methods is that SNAPP allows each branch of the species tree to have its own population size, whereas PoMo constrains all the population sizes to be equal.

A final class of methods in this broad category uses diffusion approximation to model the drift of allele frequencies within and among populations (Pickrell & Pritchard 2012). An advantage of diffusion approximation is the ability to infer migration among populations. However, the simplifying assumptions about mutation, and the conditions under which the diffusion approximation holds, limit the appropriateness of these models to very shallow timescales. As a result, the applicability of these approaches to phylogenetics is currently limited.

5.3. Quartet Methods

SVDquartets (singular value decomposition for quartets) is a quartet-based method to generate a species tree (Chifman & Kubatko 2014) (**Figure 3**). The method assumes that each SNP has an independent history given the species tree, and it is therefore considered a coalescent method similar to SNAPP (Bryant et al. 2012). Specific details of the method are provided by Chifman & Kubatko (2014) and Xu & Yang (2016). SVDquartets leverages recent advances on the algebraic geometry of site pattern probabilities on trees under continuous-time Markov chain models of sequence evolution. One advantage of this method is its statistical robustness under very general models of molecular evolution and the coalescent (Chifman & Kubatko 2015). It is also a statistically consistent estimator of the species tree, even when there is variation in evolutionary rates and effective population sizes (Long & Kubatko 2017).

SVDquartets has several advantages over other coalescent methods for estimating species trees. First, the method is fast, and it produces a species tree with estimates of uncertainty from a

bootstrapping procedure in a matter of seconds to minutes, depending on the size of the data set. Second, by subsampling quartets of species from a SNP data matrix, the problematic issue of missing data is largely circumvented, because some data will typically be present at a locus for each of the four samples included in the quartet. Finally, a study by Eaton et al. (2016) using a combination of simulation and empirical data showed that the quartet approach has the potential to provide more information for relationships deep in a phylogeny compared to those at the tips, which challenges the notion that the utility of SNP data decreases at deeper phylogenetic levels.

6. PERSPECTIVES

We predict that SNP-based phylogenies will continue to grow in popularity. The number of studies using SNPs for phylogenetic inference is growing rapidly (**Figure 1**), yet many questions remain to be addressed regarding their utility in this context. Investigators should work toward distinguishing between sources of missing data (ADO versus insufficient sequencing effort), testing the influence of modeling assumptions on phylogenetic inference, and surveying the sensitivity of results to assembly parameters. Such studies will help increase the rigor of SNP-based phylogenetic inference and should help stimulate the continued development of useful new methods. Additionally, there are currently too few comparative studies of SNPs, both empirical and in silico, versus other types of markers, or comparisons of different inference methods. More studies combining empirical data and simulation approaches will be an important part of understanding their relative benefits and biases. Finally, the development of methods that can integrate SNPs with other data types will increase their relevance in a field where integrating multiple genomic markers is an increasingly crucial part of attempting to describe the genealogy of life.

FUTURE ISSUES

Early suggestions by Rubin et al. (2012) and Cariou et al. (2013) that SNPs can provide accurate phylogenies for groups as old as 60 million years have now been put to the test across a wide variety of organisms (Leaché et al. 2015a, Eaton et al. 2016, Herrera & Shank 2016). Systematic biologists have made steady progress toward determining the limits of SNP usefulness and have identified problems and solutions for SNP phylogenetics. Many of the current challenges associated with SNP data are well characterized, and below we provide a list of areas that are in need of further investigation. We predict that SNP phylogenetics will benefit from developments across a broad array of areas that include data collection, data analysis, modeling and analysis, and data integration.

1. New data collection methods: Many current methods for collecting SNP data are efficient and cost effective, but they produce large amounts of missing data. A significant portion of nonrandom missing data is attributable to ADO (Arnold et al. 2013), which is expected to bias phylogenetic analyses (Leaché et al. 2015a). New data collection methods that combine the benefits of RADseq with those of sequence capture are promising because they appear to be more robust to ADO (Ali et al. 2016, Boucher et al. 2016, Hoffberg et al. 2016, Suchan et al. 2016). Phylogenetic studies of large clades that are likely to contain relatively ancient divergences stand to benefit greatly from these new data collection methods.

2. Extension of current methods: It is important to continue developing methods and software packages that extend existing approaches to accommodate SNP data. Phylogeographic and phylogenetic inference typically requires the estimation of gene trees (Brito & Edwards 2009). However, methods that rely on gene genealogies are not typically compatible with SNPs. Creative new solutions for SNP phylogenetics that bypass gene trees have been developed (Bryant et al. 2012, De Maio et al. 2015a), and these methods can be adapted for new applications, such as species delimitation, to further increase the utility of SNPs (Leaché et al. 2014).
3. New models: Methods that analytically integrate out gene trees (e.g., SNAPP and PoMo) are very promising, but there is room for greater generality in these models. For example, instead of reducing the data to biallelic sites, the models could accommodate the primary sequence data. Also, these methods depend on exponentiating very large rate matrices, which can slow down computation on large phylogenetic trees. Speeding up these computations will be necessary to make these methods applicable to very large trees.
4. Migration: There is a need for phylogenetic inference methods that can estimate rates of gene flow between populations or species. Phylogeographic studies in particular stand to benefit from such methods, because these investigations typically involve structured populations and/or multiple species for which reproductive isolation is incomplete. Allele frequency methods can help identify directions of gene flow among populations (Pickrell & Pritchard 2012), but the analytical machinery necessary for estimating a species tree with gene flow under a coalescent model has yet to be developed for multiple loci and species (De Maio et al. 2015b).
5. Tree space: Phylogenetic data often contain support for multiple, distinct phylogenetic trees (Sanderson et al. 2011). Careful inspections of phylogenetic tree space are necessary to investigate whether or not any subsets of the data support different phylogenetic trees. Such discrepancies can be an indicator for important biological processes, such as incomplete lineage sorting, hybridization, gene flow, or recombination. Data visualization methods, including partitioned RAD analysis (Hipp et al. 2014), are now available for exploring tree space with SNP data (Escudero et al. 2014). However, computer simulation studies are needed to help clarify the interpretation of these visualizations, because it is possible that different evolutionary processes could produce the same phylogenetic patterns.
6. Comparative studies: More detailed comparisons, using simulated and real data, between SNP trees and trees estimated from gene sequences are needed to gain a more thorough understanding of the benefits and limitations of SNPs. Additionally, a more comprehensive understanding of the impacts of data assembly parameters on SNP phylogenies would help us understand how bioinformatics decisions can potentially bias phylogenetic inferences. Ideally, a synthesis of multiple comparative SNP studies that span a wide range of divergence times could reveal some general best practices for assembling SNP data. SNPs are a promising source of data for resolving contentious phylogenetic relationships, and solving the current challenges pertaining to data structure, computational demands, and modeling assumptions is an important area for future research.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank the UW Phylogenetics Seminar and Applied Population Genetic Seminar Groups for their thoughtful comments and discussions on drafts of this review.

LITERATURE CITED

Ali OA, O'Rourke SM, Amish SJ, Meek MH, Luikart G, et al. 2016. RAD capture (Rapture): flexible and efficient sequence-based genotyping. *Genetics* 202:389–400

Allman ES, Holder MT, Rhodes JA. 2010. Estimating trees from filtered data: identifiability of models for morphological phylogenetics. *J. Theor. Biol.* 263:108–19

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* 22:3179–90

Avise JC. 2000. *Phylogeography: The History and Formation of Species*. Cambridge, MA: Harvard Univ. Press

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLOS ONE* 3:e3376

Boucher F, Casazza G, Szövényi P, Conti E. 2016. Sequence capture using RAD probes clarifies phylogenetic relationships and species boundaries in *Primula* sect. *Auricula*. *Mol. Phylogenetics Evol.* 104:60–72

Brito PH, Edwards SV. 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica* 135:439–55

Brumfield RT, Beerli P, Nickerson DA, Edwards SV. 2003. The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol. Evol.* 18:249–56

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* 29:1917–32

Cariou M, Duret L, Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. *Ecol. Evol.* 3:846–52

Chifman J, Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30:3317–24

Chifman J, Kubatko L. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *J. Theor. Biol.* 374:35–47

Chung Y, Hey J. 2017. Bayesian analysis of evolutionary divergence with genomic data under diverse demographic models. *Mol. Biol. Evol.* 34:1517–28

Collins RA, Hrbek T. 2015. An in silico comparison of reduced-representation and sequence-capture protocols for phylogenomics. bioRxiv 032565. **http://dx.doi.org/10.1101/032565**

DaCosta JM, Sorenson MD. 2014. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLOS ONE* 9:e106713

DaCosta JM, Sorenson MD. 2016. ddRAD-seq phylogenetics based on nucleotide, indel, and presence–absence polymorphisms: analyses of two avian genera with contrasting histories. *Mol. Phylogenetics Evol.* 94:122–35

De Maio N, Schlötterer C, Kosiol C. 2013. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Mol. Biol. Evol.* 30:2249–62

De Maio N, Schrempf D, Kosiol C. 2015a. PoMo: an allele frequency-based approach for species tree estimation. *Syst. Biol.* 64:1018

De Maio N, Wu CH, O'Reilly KM, Wilson D. 2015b. New routes to phylogeography: a Bayesian structured coalescent approximation. *PLOS Genet.* 11:e1005421

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24:332–40
Eaton DA, Hipp AL, González-Rodríguez A, Cavender-Bares J. 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution* 69:2587–601
Eaton DA, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Syst. Biol.* 62:689–706
Eaton DA, Spriggs EL, Park B, Donoghue MJ. 2016. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Syst. Biol.* 66:399–412
Edwards SV, Beerli P. 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* 54:1839–54
Edwards SV, Potter S, Schmitt CJ, Bragg JG, Moritz C. 2016a. Reticulation, divergence, and the phylogeography–phylogenetics continuum. *PNAS* 113:8025–32
Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, et al. 2016b. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Mol. Phylogenetics Evol.* 94:447–62
Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, et al. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *PNAS* 107:16196–200
Escudero M, Eaton DA, Hahn M, Hipp AL. 2014. Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: a case study in *Carex* (cyperaceae). *Mol. Phylogenetics Evol.* 79:359–67
Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61:717–26
Felsenstein J. 2006. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *Mol. Biol. Evol.* 23:691–700
Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, et al. 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J. Bacteriol.* 188:759–72
Garner BA, Hand BK, Amish SJ, Bernatchez L, Foster JT, et al. 2016. Genomics in conservation: case studies and bridging the gap between data and application. *Trends Ecol. Evol.* 31:81–83
Garrick RC, Bonatelli IA, Hyseni C, Morales A, Pelletier TA, et al. 2015. The evolution of phylogeographic data sets. *Mol. Ecol.* 24:1164–71
Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, et al. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* 22:3165–78
Harris RB, Carling MD, Lovette IJ. 2014. The influence of sampling design on species tree inference: a new relationship for the New World chickadees (Aves: *Poecile*). *Evolution* 68:501–13
Harvey MG, Smith BT, Glenn TC, Faircloth BC, Brumfield RT. 2016. Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Syst. Biol.* 65:910–24
Herrera S, Shank TM. 2016. RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. *Mol. Phylogenetics Evol.* 100:70–79
Hickerson M, Carstens B, Cavender-Bares J, Crandall K, Graham C, et al. 2010. Phylogeography's past, present, and future: 10 years after. *Mol. Phylogenetics Evol.* 54:291–301
Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, et al. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *PNAS* 112:12764–69
Hipp AL, Eaton DA, Cavender-Bares J, Fitzek E, Nipper R, Manos PS. 2014. A framework phylogeny of the American oak clade based on sequenced RAD data. *PLOS ONE* 9:e93975
Hoffberg SL, Kieran TJ, Catchen JM, Devault A, Faircloth BC, et al. 2016. RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Mol. Ecol. Resour.* 16:1264–78
Huang H, Knowles LL. 2016. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Syst. Biol.* 65(3):357–65
Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, et al. 2016. A new view of the tree of life. *Nat. Microbiol.* 1:16048

Hykin SM, Bi K, McGuire JA. 2015. Fixing formalin: a method to recover genomic-scale DNA sequence data from formalin-fixed museum specimens using high-throughput sequencing. *PLOS ONE* 10:e0141579

Kawakami T, Smeds L, Backström N, Husby A, Qvarnström A, et al. 2014. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Mol. Ecol.* 23:4035–58

Kuhner MK, Beerli P, Yamato J, Felsenstein J. 2000. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 156:439–47

Leaché AD, Banbury BL, Felsenstein J, Nieto-Montes de Oca A, Stamatakis A. 2015a. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Syst. Biol.* 64:1032–47

Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, Linkem CW. 2015b. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biol. Evol.* 7:706–19

Leaché AD, Fujita MK, Minin VN, Bouckaert RR. 2014. Species delimitation using genome-wide SNP data. *Syst. Biol.* 63:534–42

Leaché AD, Harris RB, Maliska ME, Linkem CW. 2013. Comparative species divergence across eight triplets of spiny lizards (*Sceloporus*) using genomic sequence data. *Genome Biol. Evol.* 5:2410–19

Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50:913–25

Linck EB, Hanna Z, Sellas A, Dumbacher JP. 2017. Evaluating hybridization capture with RAD probes as a tool for museum genomics with historical bird specimens. *Ecol. Evol.* 7:4755–67

Liu L, Xi Z, Wu S, Davis CC, Edwards SV. 2015. Estimating phylogenetic trees from genome-scale data. *Ann. N.Y. Acad. Sci.* 1360:36–53

Long C, Kubatko L. 2017. Identifiability and reconstructibility of species phylogenies under a modified coalescent. arXiv:1701.06871 [q-bio.PE]

Manthey JD, Campillo LC, Burns KJ, Moyle RG. 2016. Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid tanagers (Aves, Genus: *Piranga*). *Syst. Biol.* 65(4):640–50

McGill JR, Walkup EA, Kuhner MK. 2013. Correcting coalescent analyses for panel-based SNP ascertainment. *Genetics* 193:1185–96

McTavish EJ, Hillis DM. 2015. How do SNP ascertainment schemes and population demographics affect inferences about population history? *BMC Genom.* 16:1

Morin PA, Luikart G, Wayne RK, SNP Workshop Group. 2004. SNPs in ecology, evolution and conservation. *Trends Ecol. Evol.* 19:208–16

Nielsen R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154:931–42

Nieto-Montes de Oca A, Barley AJ, Meza-Lázaro RN, García-Vázquez UO, Zamora-Abrego JG, et al. 2017. Phylogenomics and species delimitation in the knob-scaled lizards of the genus *Xenosaurus* (Squamata: Xenosauridae) using ddRADseq data reveal a substantial underestimation of diversity. *Mol. Phylogenetics Evol.* 106:241–53

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genet.* 8:e1002967

Portik DM, Smith LL, Bi K. 2016. An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Mol. Ecol. Resour.* 16(5):1069–83

Posada D, Crandall KA. 2002. The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.* 54:396–402

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164:1645–56

Ree RH, Hipp AL. 2015. Inferring phylogenetic history from restriction site associated DNA (RADseq). In *Next-Generation Sequencing in Plant Systematics*, ed. E Hörandl, MS Appelhans, pp. 181–204. Oberreifenberg, Ger.: Koeltz Scientific

Reitzel A, Herrera S, Layden M, Martindale M, Shank T. 2013. Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Mol. Ecol.* 22:2953–70

Rubin BE, Ree RH, Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *PLOS ONE* 7:e33394

Sanderson MJ, McMahon MM, Steel M. 2011. Terraces in phylogenetic tree space. *Science* 333:448–50

Schrempf D, Minh BQ, De Maio N, von Haeseler A, Kosiol C. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. *J. Theor. Biol.* 407:362–70

Schuster SC. 2007. Next-generation sequencing transforms today's biology. *Nature* 200:16–18

Seeb J, Carvalho G, Hauser L, Naish K, Roberts S, Seeb L. 2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Mol. Ecol. Resour.* 11:1–8

Shafer AB, Wolf JB, Alves PC, Bergström L, Bruford MW, et al. 2015. Genomics and the challenging translation into conservation practice. *Trends Ecol. Evol.* 30:78–87

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Syst. Biol.* 63:83–95

Stetz JB, Sawaya MA, Ramsey AB, Amish SJ, Schwartz MK, et al. 2016. Discovery of 20,000 RAD–SNPs and development of a 52-SNP array for monitoring river otters. *Conserv. Genet. Resour.* 8:299–302

Streicher JW, Schulte JA, Wiens JJ. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. *Syst. Biol.* 65:128–45

Suchan T, Pitteloud C, Gerasimova NS, Kostikova A, Schmid S, et al. 2016. Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLOS ONE* 11:e0151651

Sutherland BJ, Gosselin T, Normandeau E, Lamothe M, Isabel N, et al. 2016. Salmonid chromosome evolution as revealed by a novel method for comparing RADseq linkage maps. *Genome Biol. Evol.* 8:3600–17

Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* 50:525–39

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, et al. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol. Ecol.* 22:787–98

Xu B, Yang Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. *Genetics* 204:1353–68

Yeates DK, Zwick A, Mikheyev AS. 2016. Museums are biobanks: unlocking the genetic potential of the three billion specimens in the world's biological collections. *Curr. Opin. Insect Sci.* 18:83–88

Zeng K, Charlesworth B. 2009. Estimating selection intensity on synonymous codon usage in a nonequilibrium population. *Genetics* 183:651–62

Zink RM, Barrowclough GF. 2008. Mitochondrial DNA under siege in avian phylogeography. *Mol. Ecol.* 17:2107–21